DOCUMENT RESUME

ED 224 812 TM 820 814

AUTHOR Hogan, Thomas P.; Mishler, Carol
TITLE Relationships Among Measures of Writing Skill.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
PUB DATE [82]
NOTE 43p.
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Testing; Criterion Referenced Tests; *Measurement Techniques; Scoring; Test Reliability; *Writing Evaluation; *Writing Research; *Writing Skills
IDENTIFIERS National Assessment of Educational Progress

ABSTRACT

This literature review summarizes what is currently known about the agreement among six measures of writing skills. Three of these methods involve the application of human judgment in scoring or rating a piece of writing: holistic, analytical, and primary trait scoring. Two methods involve anatomical or taxonomic analysis of a piece of writing: computer analysis and syntactic analysis. The final method involves the use of objective (usually multiple-choice) tests of writing-related skills. The research on relationships among the various measures of writing skills admits of relatively few well-established generalizations. Relationships among some pairs of measures have been well researched, while relationships among other pairs of measures have been virtually untouched by empirical studies. Aspect of National Assessment (NAEP) dealt with in this document: Procedures (Scoring). (Author/PN)

RELATIONSHIPS AMONG MEASURES OF WRITING SKILL

Thomas P. Hogan

&

Carol Mishler

University of Wisconsin-Green Bay

This review was prepared for the National Assessment of Educational Progress, a project of the Education Commission of the States, funded by the National Institute of Education.

Abstract

Although "writing skill" is often treated as a reasonably well-defined trait, ability, or skill, there are a variety of seemingly disparate methods all purporting to measure this skill. To what extent do these various methods agree in their measurement of writing skill? This literature review summarizes what is currently known about the agreement among six measures of writing skill: holistic, analytic, primary trait, computer-based, syntactic indices, and objective tests. Relationships among some pairs of measures have been well researched, while relationships among other pairs of measures have been virtually untouched by empirical studies.

CONTENTS

RELATIONSHIPS AMONG MEASURES OF WRITING SKILL

INTRODUCTION

There appears to be an assumption both in popular discussions of the topic as well as in the professional literature that there is such a thing as "writing skill" which is a reasonably unitary trait or ability (at least as long as we confine the reference to expository forms of writing and exclude such things as poetic writing). Despite this assumption, we have a variety of methods for assessing writing skill, some of which appear on the surface to be quite different from one another. In fact, it is not unusual to encounter authors extolling one method while condemning another, as if the different methods had nothing in common, i.e. that they were measuring radically different abilities or that one was a "good" measure and the other a "bad" measure.

One does wonder to what extent the various techniques purporting to measure writing skill are all tapping the same function. Are the distinctions among the measures merely physical and verbal, while being roughly equivalent in what they actually measure? If a curriculum program is declared successful when one technique is used as the criterion measure, is it likely that the same conclusion would have been reached had another measure been used? If it is announced to the world that students' writing skill has improved (or declined), must the announcement be qualified by a description of how the skill was assessed?

The classical problem, of course, involves the relationship between "essay tests" and "objective tests" of writing ability, a problem which is rooted in the very foundations of what we now call the field of tests and measurements and which served as a vehicle for many of the developments within that field.

However, the "essay test vs. objective test" is an oversimplified formulation of the complete question today, although still a very important segment of the question. The last dozen or so years have witnessed the emergence of several new methods purporting to measure writing skill, each being quite different in character, for one reason or another, from either essay or objective tests, as those tests have been defined traditionally. Hence, today a review of relations between alternate measures of writing skill must go much beyond the "objective and subjective testing techniques" covered by Huddleston's (1954) thorough, scholarly review of more than 25 years ago.

We have identified six types of measures currently used to assess writing skill. Three of these methods involve the application of human judgment in scoring or rating a piece of writing: holistic, analytical, and primary trait scoring. Two methods involve a kind of anatomical or taxonomic analysis of a piece of writing: computer analysis and syntactic analysis. And the final method involves the use of objective (usually multiple-choice) tests of writing-related skills. More complete descriptions of each method are provided in the next section to serve as a preface to the review of the relationships between the six measures.

(Before proceeding with the review, it may be wise to distinguish between the problem of essay vs. objective tests as measures of writing skill and the problem of essay vs. objective tests as measures of knowledge or skill in some content area such as history or mathematics. The latter issue, taken up in works such as that of Coffman (1971) is not of concern to us in this review, while the former issue is one of the central problems in our review.)

DEFINITIONS OF THE SIX MEASURES

Although each of the methods of measuring writing ability has a number of variations, each is also characterized by a basic theme or approach. We introduce the review with a description of the basic theme for each measure, with some notes on common variations and typical applications.

Holistic Scoring. In holistic scoring of essays, raters make a single, overall judgment of the quality of a piece of writing. Exactly what is meant by "quality" may vary somewhat from one study to another, but most typically it is intended to include such factors as capitalization and punctuation, aptness of word choice, grammar, organization, spelling, sentence structure, and imagination; penmanship is usually excluded. The raters are instructed to weigh all of these factors together in roughly equal proportions to form their overall judgment of quality. Raters are also instructed to make no marks (corrections, comments on the paper) and to move through each paper at a fairly rapid pace; experienced raters move through papers in the 150-300 word range (the product of 20-40 minutes of writing) at approximately a minute per paper.

The rater's final judgment is usually quantified on a point scale, ranging from low values (poor quality) to high values (high quality). There is no standard set of points to use for the scale; examples can be found of scales from 3 to 10 points.

It is highly recommended that the raters receive training in the use of the holistic method. The training is designed to ensure that raters are consistent over time and among one another; that one or two aspects of good writing are not receiving undue weight; and that the rating proceeds at an appropriate pace. Very often, the training involves the use of "anchor points," i.e. papers which have been preselected by experienced raters to illustrate various points along the score scale. By exposure to these anchor points, raters learn of the expected range of writing skill they will encounter and the degree of difference in skill represented by successive points along the score scale. Also, raters who cannot conform themselves to these anchor points after some amount of training may be eliminated from the pool of raters.

The more formal, systematic applications of holistic scoring always use trained raters. However, it must be admitted that many applications of holistic scoring have not used training, or at least it is not apparent from the description of the study whether there was any training and, if so, how much.

Most applications of holistic scoring involve the use of more than one rater per paper. There is a seemingly endless variety of ways to go beyond the use of one rater. Sometimes two raters are used, and their independent ratings are averaged or summed. (Hence, especially in the British literature, holistic scoring is sometimes referred to as "double impression" scoring; see, e.g. Wood and Quinn, 1976.). Sometimes two raters are used, and if their ratings differ by a certain number of scale points, a third rater is introduced. Sometimes each paper is read by three, four, or five raters. The practice of using two raters for each paper but introducing a third rater to resolve discrepancies seems to be gaining popularity, although by no means can it be considered the standard methodology in this area.

In addition to the usual variations on holistic methodology mentioned above, there are some unusual variations, including paired comparisons (each essay compared with each other essay), Q-sorts, rankings, and so forth. For purposes of this review, all of these will be treated as applications of the holistic method since they all follow the basic theme of making an overall judgment of the quality of writing.

The holistic method is one of the oldest procedures for assessing writing skill. For many years, it was used by the College Board, then laid to rest after much debate about its shortcomings (see Pearson, 1955), then resurrected just recently with the renewed interest in writing skill at the college level. The Hudelson English Composition Scale (Hudelson, 1921), a collection of essays representing different scale values of writing, was published in

1921, the same year (and, incidentally, by the same publisher) as the Otis Group Intelligence Scale. The preface to the test manual, extolling the merits of a systematic and direct assessment of writing skill, reads as if it were written just yesterday. One of the most noteworthy applications of holistic scoring has been its use in each of three cycles of writing assessment by the National Assessment of Educational Progress (see, e.g. NAEP, 1981). Following the practices established by NAEP, a number of states have applied holistic scoring in statewide assessments (Fredrick, 1979).

Analytic Scoring. In analytical scoring, raters score each essay on specific qualities, such as creativity, organization, mechanics, style, etc. Like holistic scoring, analytic scoring depends on subjective judgments made by raters, with variations in the number of raters used from one application to another. Sometimes the scores on the separate factors or qualities simply stand on their own, while other times the separate scores are also averaged or summed to yield a total score.

There is clearly no consensus regarding how many factors should be used in analytical scales. Examples can be found of scales with just two factors (e.g. mechanics and creativity) and of scales with as many as 18 factors. Most analytical scales, however, yield about five or six scores. One of the most well-known analytical instruments is the Composition Evaluation Scales created by Diederich, French and Carlton. Once used by the College Board, analytical scoring was discontinued, largely because the method did not prove more reliable than the more efficient holistic method. Diederich's (1974) highly readable Measuring Growth in English, often incorrectly cited as an example of holistic scoring, actually uses an analytical scale.

It is worth noting that although analytical scoring is routinely listed as one of the major approaches to the assessment of writing, actual applications of it in either the research literature or in large scale assessments (e.g.

state programs) is rather rare.

Primary Trait Scoring. In primary trait scoring, raters judge to what extent a sample of writing contains the "primary trait" that it must have in order to accomplish its purpose. Writing tasks are carefully constructed so that the purpose and audience for the piece of writing are precisely defined. Students' essays (or other written products, perhaps just a note) are then judged according to how well their writing achieves the defined purpose, i.e. exhibits the primary trait. For instance, if the dominant purpose of an exercise is explanatory, the primary trait will be explanation through selection and ordering of details. In a typical application, essays are judged by two raters on a 1-4 scale, with "1" for essays which show little or no evidence of the primary trait, "2" for essays showing minimal evidence of the primary trait, "3" for essays demonstrating competence with the primary trait and "4" for essays demonstrating excellence in the primary trait. Precisely defined scoring criteria for each score point are outlined and used for each writing task.

Essays can be rescored for secondary, tertiary, or, presumably any lower order trait. Such traits consist of any well-defined rubric for viewing the piece of writing other than the primary trait. For example, a letter, after being scored for the primary trait of whether or not the intended message was conveyed, could be scored for the secondary trait of appropriateness of letter format.

The primary trait scoring method was developed in the late '70s for the National Assessment of Educational Progress (NAEP) in response to NAEP's need to explain more fully the writing tasks that school children were able to do. It is now more prominent than holistic scoring in NAEP's writing assessments (see e.g. NAEP, 1981) and has also been adopted in many statewide writing assessments (Fredrick, 1979).

Syntactic Scoring. This approach to writing assessment is based on the analysis of grammatical forms and syntactical structures of a student's essay. Hunt's (1965) research, which revealed the ways in which children's writing becomes more syntactically complex as they advance through the grades, laid the groundwork for syntactical complexity analysis. The three major indices of syntactical complexity are words per T-unit (a subject and verb and all its surrounding modifiers), words per clause, and clauses per T-unit. In syntactical scoring, scorers segment essays into T-units and conduct other types of frequency counts of particular syntactical structures that have been shown to change as students become older. Widely used in sentence combining research, syntactical scoring has been added to the most recent NAEP writing assessment.

Syntactic analysis is very widely used by researchers whose training has been primarily in the language arts field and is infrequently used by those whose training has been in the measurement area. Hence, for example, syntactic analysis is used routinely in articles appearing in the Journal for Research in the Teaching of English, but almost never appears in the Journal of Educational Measurement.

Computer Scoring. Computer scoring of essays refers to analysis of variables within written discourse that are amenable to mechanical counts by a computer. Average word length, number and types of punctuation, sentence length, and other such features are machine counted. In this method of scoring, the essays are typed into the computer and a program to analyze countable features is run. Ordinarily, machine countable features of the writing which correlate most highly with judgments of writing quality (derived by holistic or analytical scoring) are compiled into some type of computer-generated score. Pioneering work was conducted in this area by Ellis Page (see Page, 1967, 1968; Page and Paulus, 1968), and followed up by

Slotnick (1971, 1972, 1974) and Slotnick, Knapp and Bussell (1971).

Objective Tests. A final method used to assess writing skill is provided by objective, standardized tests of the multiple-choice variety. Some of these tests, particularly ones designed for use with high school and college students, are designed specifically to assess writing skill; examples are the Test of Standard Written English, the Missouri College English Test, and the College Placement English Test. Such tests are usually formulated in terms of some logical analysis of the writing act or writing subskills, and validated in terms of the test score correlation with judgments of writing quality as represented by holistic scores on essays or grades in writing courses.

Other objective tests of language skills, particularly those included in standardized achievement batteries for use at the elementary school level, are designed to have content validity for the language arts curriculum. That curriculum, it must be noted, includes much besides writing skill. Hence, elementary school language tests often include items on library cards, types of reference works, alphabetizing, poetry, listening skills, and so forth, in addition to such presumably writing-related skills as spelling, grammar, and punctuation. Sometimes items in these different areas yield separate scores, while at other times they are simply lumped together in one Total Language score.

Some of the recent literature on the assessment of writing refers to objective tests as "indirect" measures of writing skill, while classifying such methods as holistic, analytical, and primary trait scoring as "direct" measures. The usage is unfortunate and misleading. It is true that objective tests yield an indirect measure of writing skill, but it is not true that holistic scoring (or any of the other judgment-based score) yields a direct measure of writing skill. In fact, we probably do not have direct measures of any constructs such as writing skill, or reading ability, or the myriad of

other traits, skills, and abilities of interest in educational and psychological measurement.

Other "Measures." We have identified six major methods of assessing writing skill extant in the research literature. There are, of course, an almost limitless number of other ways of looking at writing skill, including rhetorical analysis or literary criticism, error counts (spelling, subject-verb agreement, etc.), and the infamous "red-pencil-in-the-margin" of the English teacher. Error counts are sometimes included in lists of writing measures (see, e.g. "writing mechanics" in Spandel and Stiggins, 1980) and occasionally used in formal studies. But all of these "other methods" are used too infrequently in the research literature to warrant inclusion in our list of major methods of measuring writing skill.

RELIABILITY

Our interpretation of data on relationships among the six measures of writing skill, which is the main focus of attention in this paper, will be influenced by the reliability of each method. This is the classical problem of attenuation due to unreliability. Ideally, each study to be considered later would address this issue, providing information which would allow one to estimate the disattenuated relationship. Unfortunately, this is not always the case: in some studies, the relationships among measures were of only incidental concern so that the reliability issue was not explored, while in other studies the authors seem oblivious to the attenuation problem. Hence, in this section, we attempt to provide a general review of what is known about the reliability of each method, while acknowledging that these general findings may not apply to each study taken up later.

Four types of reliability determinations will enter into the discussion. First, scorer reliability will be a prominent issue for those scores which involve ratings or some other type of human judgment. Hence, scorer

reliability needs to be determined for holistic, analytical, primary trait, and syntactic maturity scores, the latter because, while some of the counts are quite mechanical, other counts do involve a judgment. For practical purposes, objective test scores and computer-based scores may be considered to have perfect scorer reliability.

There are two subcategories of scorer reliability to consider. First, there is intra-scorer reliability: the consistency with which one rater scores or judges a given set of papers on different occasions or under varying conditions. Second, there is inter-scorer reliability: the consistency with which different raters score or judge a given set of papers. Most investigations of scorer reliability deal with the latter issue.

The second major type of reliability to be determined may be referred to as alternate-form reliability. The terminology here is derived from usage within the area of objective tests, where the meaning of alternate forms is well-established. The analogous case for all of the other types of scores (all of which depend upon examinees producing a piece of writing) involves the consistency beteween scores derived from two different pieces of writing which are judged to be roughly equivalent tasks (e.g. two impromptu, 20-minute essays of an argumentative nature). In contrast, we may refer to cross-task reliability which involves consistency between scores derived from two pieces of writing which are judged to be nonequivalent tasks (e.g. writing a short thank you note vs. writing a lengthy research paper).

Finally, there are the various coefficients of internal consistency reliability applied to objective tests. Of course, from a theoretical perspective, these indices of reliability can be thought of as specific applications of alternate form reliabilities (or vice versa). They could be applied with some ease to computer and syntactic scoring, and possibly to holistic, analytic and primary trait scoring, although the application in the

latter cases might be strained beyond intelligible limits. However, as a practical matter, internal consistency reliability is used almost exclusively with objective tests and is reported separately from alternate form reliability for such tests.

True test-retest reliabilities are rarely reported for any of the measures. Even when authors do refer to test-retest reliability, they are usually using alternate form data, i.e. data based on two different writing topics.

Reliability of Holistic Scoring. One principle that has been established for a number of years is that student writing can indeed be reliably judged. Many studies have found that when proper conditions are met, interscorer reliability of .80 or above can be achieved (Cooper and Odell, 1977; Diederich, 1974; Hogan and Mishler, 1980; Page, 1968). Most researchers agree that this level of reliability is possible, despite a widespread notion to the contrary among laypersons.

On the question of alternate form reliability, opinion is somewhat more divided. Anderson (1960) notes that "the discrepancy between tests [holistically scored essays] is evidence of the unrepresentative character of a solitary essay. The significant variability among testing occasions is evidence of fluctuation in the function underlying composition ability" (p. 90). The Anderson study employed analysis of variance rather than a correlational methodology for studying reliability. Braddock, Lloyd-Jones, and Shoer (1963) cite Kincaid (1953) as also having demonstrated substantial fluctuation in writing scores across occasions, lending support to Anderson's contention that the alternate form reliability of holistically-scored essays is unacceptably low.

Hogan and Mishler (1980) report a correlation of .71 between two holistically-scored essays written on two occasions by Grade 8 students, a finding which supports Diederich's research with high school age or older

students. However, Hogan and Mishler found a slightly higher correlation of .81 at the Grade 3 level. Thus, alternate form reliability of holistic scoring appears to be noticeably lower than interscorer reliability, at least among older students.

How stable is the holistic score across writing tasks (cross-task reliability)? The topic, the mode of discourse, the time allotted for writing, the intended audience, and the instructions given to students are a few of the task variables that might presumably be investigated. Braddock et. al. (1963) cite several studies that suggest that the topic students write on influences the quality of writing produced and the resulting holistic score. Braddock et. al. suggest that mode of discourse will have a substantial effect on the holistic scores; the also notes the need for research on the optimum time needed for writing during testing. Overall, research has not been definitive on matters relating to the stability of writing across tasks, as measured by the holistic scoring method.

Reliability of Computer Scoring. In computer scoring, of course, the question of "scorer" reliability is not a problem since we are not dealing with subjective human judgments, hence, scorer reliability may be considered perfect. Page and Paulus (1968) have investigated the alternate form reliability of each of the 30 variables in their scoring system. In correlating the variables for Essay C and Essay D (written about a month apart), Page and Paulus report correlations ranging from .-02 to .65. Some of the most unreliable variables were number of slashes (.-02), presence of a title on the essay (.05), number of "Type B" declarative sentences (.09) and the number of relative pronouns (.17). Among the variables with the highest reliability were average sentence length (.63), number of commas (.61), average word length (.62), standard deviation of word length (.61), and number of common words on the Dale list (.65). Thus the alternate form reliability

of the thirty computer countable elements used in Page's study varied considerably although the variables of ultimately greatest interest (as we shall presently see) tended to have reliabilities of .60 - .65.

Reliability of Analytic Scoring. Several studies have addressed the question of interscorer reliability when analytical scoring of essays is used. Some studies have contrasted the interscorer reliability of analytical scoring with that of the faster holistic method and have come to the conclusion that the interscorer reliability of each method is about the same (Coward, 1952). A more recent investigation (Follman and Anderson, 1967) compared four analytical methods (The Diederich Rating Scale, The California Essay Scale, The Cleveland Composition Rating Scale, The Follman English Mechanics Guide) and a method similar to the holistic method, which was dubbed Everyman's Scale. Resulting average interscorer reliability coefficients ranged from .95 using the Follman English Mechanics Guide to .81 using the Cleveland Composition Rating Scale. (Reliabilities for separate subscales within the analytical scales were not reported.) Reliability using the holistic method was .95. These results show that similar levels of interscorer reliability (.80 or greater) can be attained with either holistic or analytic scoring.

In fact, the interscorer reliability coefficients reported for five different analytical scales listed in Measures for Research and Evaluation in the Language Arts (Fagan, Cooper and Jensen, 1975), a compilation of unpublished instruments, are above .80. For Diederich, French and Carlton's E.T.S. Composition Evaluation Scales, an interscorer reliability of .90 is noted. Other measures described in Measures for Research and Evaluation in the Language Arts report similarly high interscorer reliabilities of .83 for the Glazer Narrative Composition Scale, .97 for the Sager Writing Scale, .73 for the Literary Scale, and 67-100 percent agreement for the Schroeder Composition Scale.

No information concerning reliability other than scorer reliability could be found for analytic scores.

Reliability of Objective Tests. As for computer scoring, scorer reliability may be considered virtually perfect for objective tests. For regularly published objective tests, i.e. ones which have undergone the customary round of prepublication research, reliability is very much a function of test length. For objective tests with about 50 or 60 items, alternate form reliabilities are usually in the range of .85-.90 and various internal consistency measures of reliability in the range of .90-.95. For objective tests we do not really have an analog for cross-task reliability, unless entirely separate tests of English skills (say the Missiouri College English Tests vs. the College Board's Test of Standard Written English) are thought to fill this gap. Such alternate measures usually correlate about .70 -.85. Test manuals, at least for the widely used published tests, are usually chuck-full of reliability data, so we have not bothered to cite data for specific tests here.

Reliability of Syntactic Complexity Scoring. The interscorer reliability achieved when an essay is segmented into T-units consistently falls above .90. Researchers have reported that trained scorers can analyze essays for T-unit with little or no disagreement (O'Donnell, Griffin and Norris, 1967; Crowhurst and Piche, 1979). Crowhurst (1980) reported interscorer reliability coefficients ranging from .97 to .99 calculated after training and before scoring.

Alternate form reliability of the major syntactical indices (T-unit length, clause length, T-units per clause) has not been well researched. Witte and Davis (1980) have noted O'Donnell's (1976) statement: ". . . there are no data to show how consistently these indices measure the structural complexity of an individual student's writing in various situations" (p. 33).

Witte and Davis (1980), in what is apparently the only study of alternate form reliability of T-unit measures, found that T-unit length was not a stable individual trait, even within the same mode of discourse. They regard their finding as "tentative and inconclusive" and urge further research.

The stability of syntactic complexity measures across tasks has been the subject of some research that focuses on how mode of discourse influences syntactical complexity. San Jose (1972) found that mean T-unit length differed significantly across four modes of discourse. Crowhurst and Piche (1979), Crowhurst (1980) and several others have found that T-unit length produced in an argumentative essay is greater than that produced in narration. Witte and Davis (1980) also found that T-unit length was not stable across the modes of description and narration. The question of stability of T-unit length, particularly within a mode of discourse, bears further investigation. However, most of this research shows only that different tasks yield differences in average scores for syntactic measures; the issue of relative order is skirted and, hence, reliability, in the psychometric sense, is not determined.

Fredrick (1970) determined a number of syntactic indices for themes written by eighth grade students written over a six week period, then correlated the indices from the first 3-week period and second 3-week period and from odd and even pages. He found "clause length, clauses per T-unit, T-unit length, T-units per sentence, and sentence length correlated .48, .22, .56, .48, and .62, respectively, between first half and second half, and .69, .54, .74, .65, and .77 between odd and even page samples" (p. 126). It should be noted that many of the student essays used in research and in school evaluation programs are much shorter than the 1000 or 500 word samples used in this study.

Reliability of Primary Trait Scoring. Mullis (1980) reports that strong

interscorer reliability exists in primary trait scoring. Although no correlations are reported, percentages of essays on which the first and second readers agreed ranged from 91 to 96 percent for various groups of essays scored with the primary trait method (NAEP, 1981). Studies of alternate form or cross-task reliability for primary trait scores are not available.

Summary. At the risk of jeopardizing our professional reputations as well as any claims we may have to sanity, we venture the following summaries of what is presently known about the various types of reliability for each of the six methods of measuring writing skill. Table 1 indicates our judgment of how much information seems to be available regarding each type of reliability for each assessment method, while Table 2 indicates a generalized average or typical coefficient for each type of reliability for each method, at least for those instances where the amount of information allows an estimate.

Table 1. Summary of How Much Information is Available about Reliability of Each Assessment Method

Assessment Method	Scorer		Alternate Form	Cross-task	Internal Consistency
	Intra-	Inter-			
Holistic	much	much	much	little	NA[a]
Analytical	some	some	little	little	NA
Primary Trait	none	some	none	none	NA
Computer	NA	NA	some	none	none
Syntactic	much	much	little	little	little
Objective Test	NA	NA	much	much	much

[a]Not Applicable.

Table 2. Summary of Estimated Typical Reliability Coefficients for Each Assessment Method

Assessment Method	Scorer		Alternate Form	Cross-task	Internal Consistency
	Intra-	Inter-			
Holistic	.90	.85	.60	?	-
Analytical	.90	.85	.60	?	-
Primary Trait	.95	.90	?	?	-
Computer	.99	.99	.65	?	?
Syntactic	.95	.95	?	?	?
Objective Test	.99	.99	.90	.80	.90

Despite the near universal agreement about the importance of determining reliability for any measure, it seems apparent that there is still much work to be done on the reliability issue for these measures of writing skill.

RELATIONSHIPS AMONG THE MEASURES

With six different methods of measuring writing skill, we obviously have 15 possible pairings of the methods; for each pairing the question of equivalence can be raised. It is immediately apparent that some of the relationships have been studied repeatedly, while others have not been studied at all, at least as defined by the published literature. For example, the relationship between holistic scores and objective tests has been studied often, whereas the relationship of primary trait scores to any of the other methods remains a mystery. In the following sections, each relation which has been the subject of one or more studies will be treated.

Holistic vs. Computer Scoring. Page and Paulus (1968) correlated 30 computer countable variables called "proxes" with ratings of overall quality and reported a multiple correlation of .71. The proxes included such computer countable variables as average sentence length, frequency of various types of punctuation, frequency of spelling errors, standard deviation of word length

and length of essay. Page and Paulus reported moderate correlations for several of the proxes, after using the proxes to predict ratings on two separate essays. Average word length (r=.37 for essay C and .51 for essay D), standard deviation of word length (r=.45 for essay C and .53 for essay D), number of commas (r=.36 for essay C and .34 for essay D), proportion of common words on the Dale (r=-.37 for essay C and -.48 for essay D), and essay length (r=.25 for essay C and .32 for essay D) were among the best predictors of the holistic rating. Average sentence length emerged as an additional strong predictor when the reliability of each prox was taken into account.

Slotnick (1971, 1972, 1974) and Slotnick, Knapp and Bussell (1971) conducted a series of studies that built on the work of Page by expanding the computer program to include 59 indicators (in contrast to Page's 30). Vocabulary, subordination, and prepositions were computer-analyzed somewhat differently than in the Page study. In one study of college freshman writing, Slotnick et. al. (1971) report that five of the 59 indicators were significantly correlated with the holistic essay score: number of sentences (r=.379), number of logical prepositions (r=.308), number of rare words (r=.475), number of all logical prepositions, and number of quotes (r=.312). Taking the four strongest indicators together, Slotnick et. al. reported a multiple correlation of .66 between the computer-generated score and the holistic essay score. A subsequent letter writing study of two groups of adults (Slotnick, 1974) revealed the remarkably high multiple correlations of .866 and .781 when the three indicators of number of different words in the essay, mean word length, and number of misspellings were used to predict the holistic score. Thus Slotnick's overall results were similar to Page's.

Hogan and Sugano (1977) also developed a list of 30 proxes that built on the work of Page and Slotnick. They explored such proxes as vowels per word, specificity, and copulatives, in addition to the more common proxes--total

words, average word length, etc. Using 60 college freshman test essays rated holistically (high, middle and low), they obtained a multiple correlation of .65 with the proxes. Total words (r=.55), average word length (r=.20), standard deviation of word length (r=.31), number of commas per word (r=.40), and vowels per word (r=.22) were a few of the proxes that correlated positively with the holistic ratings.

Computer analysis of essays or, more precisely, computer-generated scores have yielded correlations with holistic scores in the range of .65-.86. These results, remarkably consistent across a number of different studies, seem surprisingly high; and, perhaps even more surprising is the fact that there appears to be little or no contemporary effort in this area of research.

Holistic vs. Analytic. Some research in this area has attempted to identify factors important in contributing to the holistic score. Diederich (1974) refers to the factor analysis that he, John French and Sydell Carlton conducted in 1961 on the ratings of 300 essays written by college freshmen. He identified the five factors of ideas, organization, wording, flavor, and mechanics. These factors explained 43 percent of the variance in essay scores. The holistic scoring in this study was a sorting of the essays into nine piles with no training of raters.

Few other studies as sophisticated as Diederich's exist. In Measures for Research and Evaluation in the Language Arts (Fagan et. al., 1975), a compilation of writing assessment instruments which includes many analytic scales, only one analytical instrument was validated by a correlational study. The Glazer Narrative Composition Scale (a set of 18 scales to assess the quality of young children's narrative essays) total score was found to correlate .80 with scores produced after a quick impression Q-sort. None of the 18 scale scores were individually correlated with a holistic score.

Objective Test Scores vs. Holistic Scoring. Most research investigating

the relationship of holistic essay scores to objective test scores has revealed substantial although far from perfect correlations between objective test scores and holistic ratings. Correlations generally fall in the .55-.70 range.

Research With College Students. Most research conducted on the issue of essay scores vs. objective test scores has been related to college selection and/or placement and hence, has dealt with the higher developmental levels of writing skill. The Educational Testing Service and the College Entrance Examination Board have been the major contributors to research on this question, which has been investigated fairly thoroughly with upper secondary and college students.

The widely cited study by Godshalk, Swineford, and Coffman (1966), using a largely college bound group of high school juniors and seniors, reported correlations generally in the .30's between several objective measures and single essays rated by two or three readers. But correlations of .57 to .71 were obtained between objective measures and an elaborately constructed essay score (four samples, each scored by five readers). Huddleston (1954) found a fairly high correlation between the SAT verbal subtest and instructors' ratings of student writing ability (r=.76), showing the objective test to be a better predictor of instructors' ratings than is the essay test. Pearson (1955) also reported a higher correlation between teachers' ratings of ability and the Scholastic Aptitude Test (r=.65) than between the ratings and an essay test (r=.51). Breland and Gaynor (1979) reported correlations between single essays and single scores on the objective (multiple-choice) Test of Standard Written English of .56-.63; see also Breland (1977). Similar results were reported by Wood and Quinn (1976).

Research with Elementary Children. At least three studies have researched the relationship between objective test scores and holistic essay scores among

younger children. Ondrasik, Crocker, and Lamme (1979) compared 138 fourth graders' performance on four subtests of the Metropolitan Achievement Test with their performance on two holistically scored essays, one that involved fiction-writing and one that involved a factual report task. They found moderate correlations between the holistic rating and the Word Knowledge subtest (r=.45), the Reading Comprehension subtest (r=.52), the Spelling subtest (r=.43), and the Language Arts subtest (r=.30). They concluded that the strength of the relationship observed was insufficient to suggest that standardized tests can be used to replace actual measures of writing.

Hogan and Mishler (1980) found somewhat higher correlations between Metropolitan Achievement Test subtests and holistically scored essays of third and eighth graders. They reported correlations generally in the .55-.75 range for essay scores correlated with Punctuation and Capitalization, Listening Comprehension, Usage, Grammar and Syntax, Language Study Skills, and Spelling. Correlating the total score for performance on all subtests with the holistic score produced correlations of .69-.83. Another Language subtest (part of a battery of Reading, Science, Social Studies and Math subtests) correlated .66 at grade 3 and .71 at grade 8 with holistic essay scores. Thus Hogan and Mishler found correlations of the same general magnitude as those reported in studies of college-bound students.

On the other hand, Moss, Cole and Khampalikit (1982) reported a somewhat lower correlation between the Language Test of the 3Rs Achievement Test and holistic essay scores at grade 4 (r=.20). While they reported correlations between the objective test score and the holistic essay score of .67 for grade 7 and .75 for grade 10, the lower correlation they found at grade 4 led them to conclude that "our data suggest lower relationships at the elementary school level," in contrast to the other two studies.

In sum, the relationship between objective test scores and holistically

scored essays has been reasonably well researched at the college and precollege level: correlations have revealed a substantial relationship. At least two studies have replicated these findings at the elementary level, while the other has suggested a weaker relationship for younger students' writing.

Holistic Scoring Vs. Syntactical Maturity Scoring. The relationship between quality of writing and the syntactical maturity of writing has been studied several times at the college, high school and elementary level. In general, most of these studies have found little or no relationship between quality of student writing and the syntactical complexity of the writing. Although some studies have reported gains in both syntactical maturity and quality after a particular treatment (e.g., practice in sentence-combining), these studies have not shown a high or even moderate correlation between the two measures. It should be noted that the methodology used in many studies within this category involves contrasting (then testing for significance) the syntactic indices characteristic of high-rated and low-rated essays; hence, one must often infer the strength of the relationship between holistic scores and syntactic indices from mean differences or t-values.

Research With College Students. At least two studies have reported simple correlations between college freshmen's scores on holistically scored writing samples and several of the commonly used indices of syntactic development. The sets of correlations produced in each study are remarkably low, with each study turning up almost identical correlations between quality ratings and syntactical variables. Nold and Freedman (1977) attempted to determine which of the various syntactical measures might predict the holistic scores of 22 Stanford freshmen, each of whom wrote four essays. Using the work of Golub and Fredrick (1970), Nold and Freedman correlated 17 syntactical maturity variables with quality ratings of trained raters. They found a correlation of

-.08 between words per T-unit and quality, -.09 between words per main clause and quality, -.06 between words per subordinate clause and quality, and -.03 between subordinate clauses per T-unit. (These correlations and others from the Nold and Freedman study should be read as positive correlations because a low essay score indicated high quality. Each rater used a 1-4 scale with 1 being the highest.) The variables that correlated most highly with essay quality were overall length (r=-.57), percentage of words in final free modifiers (r=-.42), percentage of finite verbs which have modal auxiliaries (r=.38) and percentage of verbs which show be or have as auxiliaries (r=.32). Nold and Freedman concluded that "words per T and other standard developmental measures are not useful in predicting perceptions of quality on the college level" (p. 174).

In studying the influence of generative rhetoric on the syntactic maturity and writing effectiveness of 138 freshmen composition students, Faigley (1979) correlated several syntactical maturity measures with holistic ratings of quality. Like Nold and Freedman, Faigley reported low correlations between quality and words per T-unit (r=.04), clauses per T-unit (r=-.07), and words per clause (r=.18). Also like Nold and Freedman, Faigley reported slightly higher correlations between quality ratings and length (r=.30), and percentage of words in final free modifers (r=.25), although the magnitude of these correlations is not quite as great as those reported by Nold and Freedman. Faigley also found a correlation of .41 between quality and percentage of T-units with final free modifiers, which was the highest correlate he reported in his study.

Gebhardt (1978) did not report correlations between quality ratings and the 86 syntactical variables used in her study of the writing of 500 freshmen. Rather, she tried to discover how quality could be measured quantitatively by determining which variables were significantly different for

33 "poor" and 21 "good" essays. She found that length of essay, mean subordinate clause length, extensive use of prepositional phrases, and coordinate conjunctive sentence beginnings were significantly different in the good and poor essays. T-unit length, on the other hand, was not significantly different. Martin (1980) found no relationship between T-unit length and ratings of freshman essays; rather, clause endings, free modifiers, and percentage of common verbs were significantly related to high quality of writing.

At the college level, then, the evidence suggests that the relationship between commonly used syntactical maturity measures and quality ratings is generally weak.

Research With Students in Grades 2-12. Early developmental research, such as that of Hunt (1965), Bateman and Zidonis (1966) and O'Donnell, Griffin and Norris (1967), did not generally concern itself with the relationship between quality of writing and syntactical measures. As Hunt said about his 1965 landmark study of grammatical structures at three grade levels, "In this study the word 'maturity' is intended to designate nothing more than 'the observed characteristics of writers in an older grade.' It has nothing to do with whether older students write 'better' in any general stylistic sense" (p. 5).

However, in addition to measuring syntactical growth after a particular course of study (e.g., transformational grammar instruction), some researchers measured the quality of students writing as a kind of secondary post-test. Mellon (1969) and O'Hare (1973) both included quality ratings in their experimental studies of transformational grammar and sentence combining, respectively, but neither reported correlations between the two measures. Mellon (1969) found that judged quality of writing actually decreased among the experimental groups, while syntactical maturity increased among the experimental groups who had undergone transformational grammar study. O'Hare

(1973) found both quality of writing and syntactical maturity increased in the experimental groups who had practiced sentence-combining. Sullivan (1978) found that sentence combining exercises did enhance syntactic maturity but did not have an effect on overall quality of writing of eleventh grade students. Callaghan (1978) reported a similar conclusion for ninth grade students.

Several studies at the elementary and high school levels have more directly investigated the relationship between various syntactic measures and quality of writing by use of a contrasted group methodology. Golub and Fredrick (1970), in their study of the linguistic structures and deviations of writing of 160 fourth and sixth graders, compared high, middle, and low rated essays on 63 measures of linguistic structure. They found that many linguistic variables were significantly different for the high and low rated essays, but words per T-unit, clauses per T-unit, and words per clause were not among the significant variables; see also Golub and Fredrick (1971). Jurgens and Griffin (1970) found little relationship between overall quality and seven language features in compositions written in grades seven, nine, and eleven. They, like Golub and Fredrick, did not report correlations between quality ratings and syntactical measures. Stokes (1979) found no significant relationship between quality of writing and T-unit length in the writing of eighth, tenth, and twelfth graders, nor did Evans and Perkins (1979) in their analysis of fourth, eighth and eleventh graders in the Oregon Statewide Writing Assessment.

Veal (1974) studied the relationship between holistic scores and syntactical measures "as a validity study" for syntactic measures. Although he did not correlate the measures directly, he found that syntactic measures clearly distinguished between high and low quality writing in the second, fourth, and sixth grades. More specifically, he found that words per T-unit distinguished between high and low rated essays at all three grade levels, but within some grade levels it failed to distinguish between high and middle

essays or between low and middle or some other combination other than low vs. high. Hence, from this study one would infer a significant but weak relationship between rated quality and T-unit length.

Several studies report a significant relationship between syntactical variables and overall quality ratings. None report correlations. Chew (1978), in an analysis of 57 New York Regents essays, found that the papers with the longest T-units were among those receiving the highest grades. Dilworth, Reising and Wolfe (1978) found that superior-rated high school essays contained more words per T-unit, were longer, and exhibited higher levels of abstraction than lower rated essays. Likewise, Distefano and Marzano (1978), in their analysis of 450 NAEP essays, found that T-unit length was a significant factor for predicting holistic scores for 9 year olds and 13 year olds, but not for 17 year olds.

Crowhurst (1980) suggested that mode of discourse could significantly influence the relationship between quality and syntactic complexity. She found that high syntactic complexity was not associated with high quality ratings if the mode was narration. However, in argumentative writing, high syntactic complexity was associated with higher quality ratings at grades 10 and 12 but not at grade 6.

At least two studies at the high school or elementary level have directly correlated essay scores with syntactical measures and both have found similar, low correlations. Howerton, Jacobson, and Seldon (1977) correlated Composition Evaluation Scale essay scores[1] with words per T-unit and reported correlations of .17 at grade four, .13 at grade six, .31 at grade

[1]This score is generated through an analytical scale, in which the essay is judged on eight factors. However, since the Howerton et. al. study did not correlate each scale score with the syntactical measures, discussion of this study seems to fit under "holistic vs. syntactic." Although the rating method was not holistic, a single score was produced to indicate quality.

nine, and .18 at grade twelve. These correlations were not as high as those found between overall length and quality (r=.30 to .54) and between percentage of total words misspelled (r=-.27 to -.50). The conclusion reached was that qualitative and quantitative measures are related since their stepwise multiple regression showed that from 21% to 57% of the variance between quality ratings can be accounted for using the five variables of total length, total sentences, percent of unique words written, percent of unique words misspelled, and words per T-unit. However, only one of the common syntactic variables, words per T-unit, was used in this study and, as shown, it did not correlate highly with quality ratings.

Stewart and Grobe (1979) investigated the relationship between fifth, eighth, and eleventh grade students' syntactical maturity and quality ratings given by trained teachers. In contrast to the Howerton, Jacobson and Selden study, Stewart and Grobe correlated quality ratings with words per clause and clauses per T-unit, as well as words per T-unit and some others. They reported signficant correlations between quality of writing and words per T-unit (r=.30), words per clause (r=.23), and clauses per T-unit (r=.37) at grade five only. For grades eight and eleven, lower correlations were reported--for words per T-unit vs. quality at grade 8 (r=.19) and for words per T-unit vs. quality at grade eleven (r=-.06). The correlations between quality and words per clause and clauses per T-unit fell into the similarly low range of -.19 to .20. Stewart and Grobe concluded that no strong significant relationship exists between holistic scores and any of the three common measures of syntactic development, except at the grade 5 level. They also concluded, as others have, that overall length correlates more highly with quality (r=.36-.47) than do the syntactical measures. Grobe's (1981) more recent study, a stepwise multiple regression, showed that none of 14 syntactical variables by themselves could accurately predict holistic scores

at grades 5, 8, or 11.

Several studies, then, have established at both the college level and lower levels that measures of syntactical development seem to bear, at best, weak relationships to the rated quality of writing.

Objective Tests vs. Syntactical Complexity Measures. In most research on the relationship between syntactical complexity and writing quality, rated essays are used as correlates or as criterion measures in the prediction of quality. Since objective tests of writing skills are widely used to measure writing and language growth, the relationship between these objective measures and the major indices of syntactical complexity would seem to be important. To what extent do T-unit counts, for example, correlate with particular objective language test or subtest scores? The relationship between syntactic measures and objective language tests has not been well researched.

Simpson (1974) conducted a canonical and multiple correlation study of measures of writing of 402 fourth, fifth, and sixth graders. Instead of attempting to predict quality ratings, Simpson identified significant predictors of two objective test scores and an essay score, using the language portion of the Iowa Test of Basic Skills, the Watts Test of Connecting Words and Phrases, and the Writing Test (an essay test) of the Sequential Test of Educational Progress. Student writing samples were scored for 56 predictor measures, including words per T-unit. He found the Myklebust syntax score, a weighted ratio of errors to words written, to be the most important predictor of objective test performance with canonical correlations in the neighborhood of .83 or above. T-unit length alone did not emerge as an important predictor, leading Simpson to conclude that "attempts to classify children or evaluate English programs solely on measures of T-unit length and transformational structures do not account for the major factors of writing ability."

Ondrasik, Crocker, and Lamme (1979) also completed a canonical correlation study of the relationship between four objective subtest scores and measures of writing proficiency. However, neither words per T-unit nor any of the other common syntactical indices were used in the analysis. Rather, total number of T-units was used as a variable. Low correlations of .17 and -.02 were reported between number of T-units and performance on the objective subtests.

No other studies comparing performance on objective language tests with syntactical complexity of writing samples could be found.

Computer vs. Analytic. Page and Paulus (1968), in addition to their work on computer prediction of holistic essay scores, also examined the relationship of their thirty proxes to five analytical ratings. The analytical scale included separate ratings of essays for creativity, ideas, style, mechanics, and organization. The correlations were all in the moderate to high range: creativity (r=.78), ideas (r=.78), style (r=.77), mechanics (r=.64) and organization (r=.69). The surprising finding that a composite of the 30 proxes was correlated most highly with creativity ratings seems to be accounted for in large measure by the contribution of the "essay length" prox. For the average of all five traits vs. the thirty proxes, Page and Paulus report a multiple correlation of .72, similar to the multiple correlation found between the holistic scores and the proxes. Those proxes contributing the most to the prediction of the average of the five traits were length (r=.26), commas (r=.38), dashes (r=.32), standard deviation of word length (r=.45) and spelling errors (r=-.19).

Syntactic vs. Computer. Golub and Kidder (1974) have developed the Syntactical Density Score (SDS) which uses computer analysis of essays to produce a measure of syntactic maturity; see also Golub (1974). The SDS was designed by selecting the best 10 of 63 variables that attempted to predict

quality of writing in Golub and Fredrick's 1970 and 1971 studies, discussed elsewhere in this paper. The ten variables are: 1) words per T unit; 2) subordinate clauses per T unit; 3) main clause word length; 4) subordinate clause word length; 5) number of modals; 6) number of be and have forms in the auxiliary position;
7) number of prepositional phrases; 8) number of possessives; 9) number of adverbs of time; 10) number of gerunds, participles, and absolute phrases (unbound modifiers).

The computer program makes "decisions" about the syntactic structures that "probably" exist due to the pattern of punctuation in the essay. Kidder and Golub (1974) report a correlation of .96 between computer generated and hand tabulated scores for the syntactic features.

Analytic vs. Objective Test. Few studies compare performance on objective tests with analytically rated characteristics of student papers. Usually, the overall score produced through analytical rating would be correlated with some criterion (e.g., Howerton et. al., 1977) but rarely are ratings on particular traits correlated with a criterion such as an objective test score.

DISCUSSION AND GENERALIZATIONS

The research on relationships among the various measures of writing skill admits of relatively few well-established generalizations. Nonetheless, in this final section we attempt to formulate a number of conclusions, identify major questions yet to be answered, and discuss some other problems relevant to the measurement of writing skill.

1. The relationship between holistic ratings of essays and objective test scores has been fairly well established. Correlations between the two types of measures are generally about .60. If this figure is corrected for unreliability in the objective test and in the scoring of the essay, the r increases to about .70, but if the correction is made to

include the alternate form or cross-task reliability of the essay, the corrected r would be in the neighborhood of .80 or better.

Recent research on the relationship between holistic scores and objective tests differs little either in its methodology or conclusions from that summarized by Huddleston (1954). It might also be noted that there has been no abatement over the years in the disbelief in, even outright rejection of, these findings.

2. Although the research on scorer reliability is now quite clear, i.e. essays can be scored quite reliably, the reliability of essays across occasions or types of tasks has not been thoroughly documented. Evidence available on this latter issue, although meager, suggests the presence of a disconcerting amount of unreliable variance across occasions and tasks; and this problem would seem to beset all of the methods which depend upon a writing sample, i.e. all methods except objective tests.

3. While analytic scales are invariably listed among the various methods of measuring writing skill, they are used very little in the formal research literature (and perhaps anywhere else, too). The bits of evidence which we do have about scores derived from analytical scales suggest that they behave very much like holistic scores, both in terms of the subscores and, even more so, in terms of the frequently used total score obtained from analytical devices. In other words, the subscales contain little unique variance, certainly far less than the originators and proponents of analytic scales suppose. Hence, for practical purposes, it is probably safe to assume that any generalizations developed for holistic scores will hold true for analytic scales, too.

4. Various syntactical measures bear little relationship to holistic ratings of quality of writing (and, therefore, presumably to analytical ratings) or to objective test scores. The relationships tend to be

negligible or, if significant at all, very weak. One does begin to wonder what the syntactic indices are measuring. To be sure, some authors state quite clearly that syntactic indices are intended to simply describe language, not to measure its quality. But it is important to note that the syntactic indices are often used in practice to recommend continuation (or discontinuation) of instructional strategies and programs which apparently are designed to improve the quality of writing.

5. Computer generated scores (weighted composites of computer countable features of a written work) yield surprisingly high correlations with the quality of writing, as defined by holistic scores. The correlations are generally in the range of .60-.70. Even some of the individual computer-counted features, such as length of essay, mean and standard deviation of word length, and indices of vocabulary load, consistently yield significant though moderate correlations with rated quality. Strangely, however, no research on computer generated scores has been published since the spurt of activity with this method in the late '60's and early '70's.

6. It seems odd that the two latter generalizations (#'s 4 and 5) could be simultaneously true, since computer analysis and syntactic analysis seem to have so much in common. Sometimes it almost seems as if the syntactic analysis is too sophisticated, laying ever more complexly and obscurely defined indices on top of one another, thereby missing what are perhaps some rather simple, direct qualities of good writing. To be sure, that explanation, if not downright philistine, is at least not very helpful. Or, it may be that the success of the computer generated score lies mainly in its reliance on combining several variables, each of which has rather limited reliability, whereas the syntactic indices, each with rather limited reliability, usually stand alone. In any case, this question seems to beg for further analysis.

7. Primary trait scoring has been the subject of virtually no published research. It hardly seems appropriate to foist this method upon the world at this time, although evidently it is being pedalled across the country in an almost cavalier fashion. We know practically nothing about the measurement characteristics of the primary trait method of scoring: its reliability as defined in the usual variety of ways, its relationship to other measures, its relationship to external criteria, etc. Because it seems like a good idea hardly seems like an adequate basis for widespread, routine use of the technique, at least if we pay any respects at all to fundamental notions of good measurement practice. All of this is not to say that primary trait scoring is not a good measurement technique. It is only to say that at the present time we don't know very much about its measurement characteristics and, therefore, ought to confine its use to restricted research applications.

8. There are a number of issues lurking in the literature on writing assessment which fairly cry out for empirical analysis. Without pretending to draw up an exhaustive list of these, we offer the following three topics as being high priority items in any research agenda. The first, which has already been mentioned, is the cross-task generalizability of the various types of scores derived from writing samples. There is a widespread feeling that different types of tasks, as defined, for example, by the traditional "modes of discourse" (argumentative, narrative, etc.) yield different results. Indeed, there is now good evidence that certain features of writing differ from one of these types of tasks to another. But these are average differences and may not affect relative order of performance; that is, the differences discovered to date may be nothing more than scale transformations. We simply don't know.

A second issue relates to the length of writing sample required for analysis. One finds rather strongly propounded opinions on this point, with recommendations ranging from 20 minutes to two hours. However, there appears to be no empirical evidence on this issue.

Finally, while it is generally accepted that training of raters is an important prerequisite for use of scoring methods which depend heavily on human judgment, there seems to be no evidence regarding how much training is enough. In many practical applications, training may be rather lengthy. In other instances, training is brief in the extreme, consisting of reading a page of instructions and having a 5-minute discussion. Our suspicion is that some of the more elaborately designed training sessions are more fluff than substance, intended more for public relations than reliability. However, the issue is empirically resolvable and really should be addressed by a number of studies.

REFERENCES

Anderson, C. C. The new STEP essay test as a measure of composition ability, Educational and Psychological Measurement, 1960, XX, 95-102.

Bateman, D. and F. Zidonis. The effect of a study of transformational grammar on the writing of ninth and tenth graders. Research Report No. 6. Urbana, IL: National Council of Teachers of English, 1966.

Braddock, R., R. Lloyd-Jones, and L. Shoer. Research in Written Composition. Urbana, IL: National Council of Teachers of English, 1963.

Breland, H. and J. Gaynor. A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 1979, 16, 119-127.

Breland, H. M. A study of college English placement and the Test of Standard Written English. Princeton, NJ: Educational Testing Service, 1977.

Callaghan, T. F. The effects of sentence combining exercises on the syntactic maturity, quality of writing, reading ability and attitudes of ninth grade students. Dissertation Abstracts International 1978, 39, 637A.

Chew, C. A study to determine the relationship between indexes of maturity and a quality grade in a written examination. Educational Resources Information Center, 1978. (ERIC Document Reproduction Service No. ED 121 799).

Coffman, W. E. Essay examinations. In R. L. Thorndike (Ed.) Educational Measurement (2nd Ed.). Washington, D. C.: American Council on Education, 1971, pp. 271-302.

Cooper, C. and L. Odell. Evaluating Writing: Describing, Measuring, Judging. Urbana, IL: National Council of Teachers of English, 1977.

Coward, A. F. A comparison of two methods of grading English compositions. Journal of Educational Research, 1952, 46, 81-93.

Crowhurst, M. and G. Piche. Audience and mode of discourse effects on syntactic complexity in writing at two grade levels. Research in the Teaching of English, 1979, 13, 101-109.

Crowhurst, M. Syntactic complexity and teachers' quality ratings of narrations and arguments. Research in the Teaching of English, 1980, 14, 223-231.

Diederich, P. B. Measuring Growth in English. Urbana, IL: National Council of Teachers of English, 1974.

Dilworth, C. B., Jr., R. W. Reising and D. Wolfe. Language structure, and thought in written composition: Certain relationships. Research in the Teaching of English, 1978, 12, 97-106.

DiStefano, P. and R. Marzano. Skills in composition: A new approach English Education, 1978, 9, 117-121.

Evans, D. and Perkins. A syntactic analysis of written compositions from 1978 Oregon statewide assessment. Paper prepared for the Oregon Department of Education, June 1979.

Fagan, W. T., C. R. Cooper and J. M. Jensen. Measures for Research and Evaluation in the English Language Arts. Urbana, IL: National Council of Teachers of English, 1975.

Faigley, L. The influence of generative rhetoric on the syntactic maturity and writing effectiveness of college freshmen. Research in the Teaching of English, 1979, 13, 197-205.

Follman, J. C. and J. A. Anderson. An investigation of the reliability of five procedures for grading English themes. Research in the Teaching of English, 1967, 2, 190-200.

Fredrick, V. Writing Assessment Research Report: A National Survey. Madison, WI: Wisconsin Department of Public Instruction, Pupil Assessment Program, 1979.

Fredrick, W. C. Reliability of measures from 500 written words. Psychological Reports, 1970, 27, 126.

Gebhard, A. Writing quality and syntax: A transformational analysis of three prose samples. Research in the Teaching of English, 1978, 12, 211-231.

Godshalk, F. I., F. Swineford and W. E. Coffman. The Measurement of Writing Ability. New York: College Entrance Examination Board, 1966.

Golub, L. S. Syntactic density score (SDS) with some aids for tabulating. Educational Resources Information Center, 1974. (ERIC Document Reproduction Service No. ED 091 741).

Golub, L. & W. Fredrick. Linguistic structures and deviations in children's written sentences. Technical Report from the Wisconsin Research and Development Center for Cognitive Learning. The University of Wisconsin, No. 152, 1970.

Golub, L. and W. Fredrick. Linguistic structures in the discourse of fourth and sixth graders. Technical Report from the Wisconsin Research and Development Center for Cognitive Learning. The University of Wisconsin, No. 166, 1971.

Golub, L. S. and C. Kidder. Syntactic density and the computer. Elementary English, 1974, 51, 1128-31.

Grobe, C. Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. Research in the Teaching of English, 1981, 15, 75-85.

Hogan, T. and C. Mishler. Relationships between essay tests and objective tests of language skills for elementary school students. Journal of Educational Measurement, 1980, 17, 219-227.

Hogan, T. and N. Sugano. The freshman essay test: Multiple approaches to its analysis and use. Presentation at the Annual Meeting of the Wisconsin Educational Research Assoc., Oshkosh, WI, December 1977.

Howerton, M. L., M. Jacobson, and R. Selden. The relationship between quantitative and qualitative measures of writing skills. Paper presented to the American Educational Research Association, New York, 1977.

Huddleston, E. Measurement of writing ability at the college-entrance level: Objective vs. subjective techniques. Journal of Experimental Education, 1954, 22, 165-213.

Hudelson, E. Hudelson English Composition Scale. New York: World Book Company, 1921.

Hunt, K. W. Grammatical structures written at three grade levels: Research Report No. 3. Urbana, IL: National Council of Teachers of English, 1965.

Jurgens, J. M. and W. J. Griffin. Relationships between overall quality and seven language features in compositions written in grades seven, nine, and eleven. Nashville, Tenn.: Institute of School Learning and Individual Differences, George Peabody College for Teachers, 1970.

Kidder, C. L. and L. S. Golub. Computer application of a syntactic density measure. Educational Resources Information Center, 1974. (ERIC Document Reproduction Service No. ED 090 304.)

Kincaid, G. L. Some factors affecting variations in the quality of students' writing. Unpublished doctoral dissertation. Michigan State University, 1953.

Martin, C. A. Syntax and success: Stylistic features of superior freshman essays. Dissertation Abstracts International, 1980, 40, 4010A.

Mellon, J. Transformational sentence combining: A method for enhancing the development of syntactic fluency in English composition. Research Report No. 10. Urbana, IL: National Council of Teachers of English, 1969.

Moss, P., N. Cole, and C. Khampalikit. A comparison of procedures to assess written language skills at grades 4, 7, and 10. Journal of Educational Measurement, 1982, 19, 37-47.

Mullis, I. V. S. Using the primary trait system for evaluating writing. Report No. 10-W-51. Denver, Colorado: National Assessment of Educational Progress, 1980.

National Assessment of Educational Progress. Procedural Handbook, 1978-79 Writing Assessment. Denver: Education Commission of the States, 1981.

Nold, E. W. and S. W. Freedman. An analysis of readers' responses to essays. Research in the Teaching of English, 1977, 11, 164-174.

Noyes, E. S. Essay and objective tests in English. College Board Review, 1963, 49, 7-10.

O'Donnell, R. C. A critique of some indices of syntactic maturity. Research in the Teaching of English, 1976, 10, 31-38.

O'Donnell, R., W. Griffin and R. Norris. Syntax of kindergarten and elementary school children: A transformational analysis. Research Report No. 8. Urbana, IL: National Council of Teachers of English, 1967.

O'Hare, F. Sentence combining: Improving student writing without formal grammar instruction. Research Report No. 15. Urbana, IL: National Council of Teachers of English, 1973.

Ondrasik, T., L. Crocker, and L. Lamme. Predicting children's writing performance from standardized achievement tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April 1979.

Page, E. and D. Paulus. The analysis of essays by computer. U.S. Department of Health, Education and Welfare, Final Report of Project No. 6-1318, April 1968.

Page, E. B. Grading essays by computer: Progress report. In Proceedings of the 1966 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1967, pp. 87-100.

Page, E. B. The use of the computer in analyzing student essays. International Review of Education, 1968, 210-225.

Pearson, R. Should the general composition test be continued? The test fails as an entrance examination. College Board Review, 1955, 25, 2-9.

San Jose, C. Grammatical structures in four modes of writing at the fourth grade level. Unpublished doctoral dissertation. Syracuse University, 1972.

Simpson, G. F. Measures of writing ability of fourth, fifth and sixth grade children. Dissertation Abstracts International, 34, 1974, 5497A.

Slotnick, H. B. Computer scoring of formal letters. Journal of Business Communications, 1974, 11, 11-20.

Slotnick, H. B. Toward a theory of computer essay grading. Journal of Educational Measurement, 1972, 9, 253-263.

Slotnick, H. B. An examination of computer grading of essays. Unpublished doctoral dissertation. University of Illinois, 1971.

Slotnick, H. B., J. V. Knapp and R. L. Bussell. Bits, nybbles, and bytes: A view of an electronic clerk. Journal of Business Communications, 1971, 8, 35-52.

Spandel, V. and R. J. Stiggins. Direct Measures of Writing Skill: Issues and Applications. Portland, OR: Northwest Regional Educational Laboratory, 1980.

Steward, M. and C. Grobe. Syntactic maturity, mechanics of writing and teachers' quality ratings. Research in the Teaching of English, 1979, 13, 207-215.

Stokes, P. W. The quality of student composition as predicted by average number of words per T-unit and organization skills. Dissertation Abstracts International, 1979, 39, 4036A-37A.

Sullivan, M. A. The effects of sentence combining exercises on syntactic maturity, quality of writing, reading ability and attitudes of students in grade eleven. Dissertation Abstracts International, 1978, 39, 1197A.

Veal, L. R. Syntactic measures and rated quality in the writing of young children. Studies in Language Education, Report No. 8. Athens: University of Georgia, 1974.

Witte, S. and A. Davis. The stability of T-unit length: A preliminary investigation. Research in the Teaching of English, 1980, 14, 5-17.

Wood, R. and B. Quinn. Double impression marking of English language essay and summary questions. Educational Review, 28, 1976, 229-46.